BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations

Yves Dehouck*, Jean Marc Kwasigroch, Marianne Rooman and Dimitri Gilis

Department of BioModelling, BioInformatics and BioProcesses, Université Libre de Bruxelles (ULB), CP165/61, Av. Fr. Roosevelt 50, 1050 Brussels, Belgium

ABSTRACT

The ability of proteins to establish highly selective interactions with a variety of (macro)molecular partners is a crucial prerequisite to the realization of their biological functions. The availability of computational tools to evaluate the impact of mutations on protein–protein binding can therefore be valuable in a wide range of industrial and biomedical applications, and help rationalize the consequences of non-synonymous single-nucleotide polymorphisms. BeAtMuSiC (http://babylone.ulb.ac.be/beatmusic) is a coarse-grained predictor of the changes in binding free energy induced by point mutations. It relies on a set of statistical potentials derived from known protein structures, and combines the effect of the mutation on the strength of the interactions at the interface, and on the overall stability of the complex. The BeAtMuSiC server requires as input the structure of the protein–protein complex, and gives the possibility to assess rapidly all possible mutations in a protein chain or at the interface, with predictive performances that are in line with the best current methodologies.

INTRODUCTION

The formation of protein complexes plays an essential role in the regulation of numerous biological processes. The rational design or modification of the affinity and specificity of protein–protein interactions is a challenging issue that stimulated considerable efforts, as it presents many promising applications, notably for therapeutical purposes (1,2).

The characteristics of protein interfaces have been thoroughly investigated (3–10). Even if the diversity of binding modes precludes the identification of a simple set of general rules, a number of common features have been underlined, such as the importance of hydrophobic contacts and electrostatic interactions at the interface. Importantly, it has also been shown that a small fraction of the residues participating to the protein–protein interface are generally responsible for most of the binding affinity (11–13). These critical residues, commonly referred to as 'hotspots', are usually defined as positions where a mutation would cause an increase of the binding free energy of at least 2.0 kcal/mol. Alanine scanning mutagenesis has been widely used to experimentally characterize protein–protein interfaces and identify these hotspots, which constitute prime targets for the modulation of protein–protein interactions (14,15).

Considerable attention has been devoted to the development of computational methods for the identification of hotspot residues in protein–protein interfaces (16–29). Most rely on a machine learning technique to integrate a variety of features characterizing each residue and its environment. These features typically include information about sequence conservation, as well as physicochemical (e.g. residue hydrophobicity, electrostatic charge), structural (e.g. solvent accessibility, number of contacts, secondary structure), or energetic parameters. Although knowledge of the structure of the complex is generally required, methods have also been implemented to predict the localization of hotspots directly from the sequence (18), or from docking simulations (22).

Besides the binary classification of hotspot residues, a more general challenge consists in the estimation of the impact of mutations on the free energy of binding. Molecular mechanics combined to continuum solvent models, MM-PBSA or MM-GBSA (MM: molecular mechanics, PB: Poisson-Boltzmann, GB: generalized Born, SA: surface area), have been exploited for that purpose (30–33). Less computationally intensive approaches, based on empirical energy functions coupled with a somewhat simplified representation, have also been described (34–37). With a few exceptions (34,37), these methods have so far been mainly focused on evaluating the effects of mutations into alanine, but not into other types of amino acids.

We present here a webserver for the prediction of changes in protein–protein binding affinity on mutations. BeAtMuSiC is based on a set of statistical potentials adapted to a coarse-grained representation of protein

---

*To whom correspondence should be addressed. Tel: +32 2 6503615; Fax: +32 2 6503575; Email: ydehouck@ulb.ac.be

structures, which allows the fast assessment of all possible mutations in the protein complex. Originally parametrized on the basis of a data set of mutations into alanine (38), our approach is here validated on a much larger data set including mutations into any kind of amino acid (39). In addition, our method stood among the top performers during the 26th round of the blind prediction experiment Critical Assessment of PRedicted Interactions (CAPRI) (40), which consisted in the evaluation of ~2000 mutations in two designed inhibitors of influenza hemagglutinin.

METHODOLOGY

Binding models

Two different binding models are considered (38). In the first model, both partners of the interaction are assumed to be able to fold independently of each other. The change in binding free energy ($\Delta\Delta G_{B'}$) resulting from a mutation is then expressed as follows:

$$\Delta\Delta G_{B'} = \Delta\Delta G_C - (\Delta\Delta G_{P_1} + \Delta\Delta G_{P_2}) \qquad (1)$$

where $\Delta\Delta G = \Delta G^{\text{mutant}} - \Delta G^{\text{wildtype}}$, ΔG_{P_1} and ΔG_{P_2} are the respective folding free energies of the two partners, and ΔG_C is the folding free energy of the complex as a whole (Figure 1). In the second model, the partners are unable to fold independently, and the change in binding free energy on mutation is thus given by

$$\Delta\Delta G_{B''} = \Delta\Delta G_C \qquad (2)$$
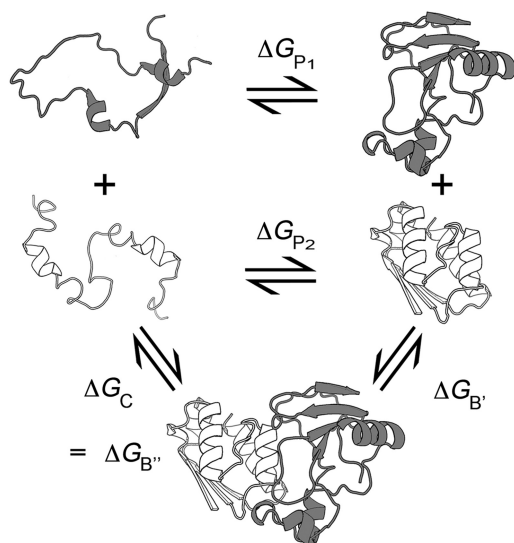
Figure 1. Schematic representation of the binding and folding free energies. ΔG_{P_1} and ΔG_{P_2} are the folding free energies of the two partners of the interaction. ΔG_C is the folding free energy of the complex as a whole. In the first binding model, the complex is formed from the association of two individually folded partners, and the binding free energy is $\Delta G_{B'}$. In the second binding model, the proteins are unable to fold independently, and the binding free energy ($\Delta G_{B''}$) is thus equal to the folding free energy of the complex (ΔG_C). The figure was made using PyMOL.

Change in folding free energy on mutation

$\Delta\Delta G_{B'}$ and $\Delta\Delta G_{B''}$ can be obtained from the estimation of the impact of the mutation on the folding free energy of the partners ($\Delta\Delta G_{P_1}$ and $\Delta\Delta G_{P_2}$) and of the complex ($\Delta\Delta G_C$). These contributions are computed as follows:

$$\Delta\Delta G = \sum_{i=1}^{13} \alpha_i(\mathcal{A})\Delta\Delta W_i \\ + \alpha_{14}(\mathcal{A})\Delta V_+ + \alpha_{15}(\mathcal{A})\Delta V_- + \alpha_{16}(\mathcal{A}) \qquad (3)$$

where $\Delta\Delta W_i$ corresponds to the energetic change induced by the mutation, according to 1 of 13 different statistical potentials extracted from a data set of known protein structures (41). These potentials describe the correlations between amino acid types, pairwise inter-residue distances, backbone torsion angles and solvent accessibilities. The inter-residue distances are evaluated from the coordinates of the average geometric centers of the side chains. In addition, the terms ΔV_+ and ΔV_- were introduced to account for the possible creation of packing defects: $\Delta V_\pm = (V_m - V_w)\mathcal{H}[\pm(V_m - V_w)]$, where $\mathcal{H}$ is the Heaviside function, and V_w and V_m the volume of the wild-type and mutant amino acid, respectively. The weights $\alpha_i(\mathcal{A})$ ($i = 1, \ldots, 16$) are sigmoid functions of the solvent accessibility $\mathcal{A}$ of the mutated residue, and were identified on the basis of a data set of experimentally measured changes in folding free energy resulting from 2648 mutations (42).

Change in binding free energy on mutation

Although the behavior of many interacting proteins is probably better described by the first binding model [Equation (1)], numerous examples of natively unfolded proteins—or protein regions—that fold only on binding have also been reported (43). In addition, even if the two partners adopt well-defined structures individually, their interaction may in some cases induce extensive conformational rearrangements. Given the limited amount of currently available experimental mutagenesis data, it is, however, difficult to build prediction tools that account for the singular properties of each protein–protein interaction. The predictions of BeAtMuSiC are thus based on the assumption of an intermediate situation, and the change in binding affinity on mutation is obtained by combining the output of the two binding models:

$$\Delta\Delta G_B = v(w\Delta\Delta G_{B'} + (1 - w)\Delta\Delta G_{B''}) \qquad (4)$$

where v and w are adjustable parameters set to 1.25 and 0.7, respectively, after optimization on a data set of 362 mutations into alanine (35,38). The Pearson correlation coefficient between measured and predicted $\Delta\Delta G_B$ values reached 0.55 on the full data set and 0.76 on 90% of the data set (Table 1).

Because conformational rearrangements on binding are not explicitly modeled, $\Delta\Delta G_C$ and $(\Delta\Delta G_{P_1} + \Delta\Delta G_{P_2})$ are rigorously equal for mutations outside of the interface region. $\Delta\Delta G_{B'}$ is thus focused on the impact of the mutation on the interactions established at the interface [Equation (1)], whereas $\Delta\Delta G_{B''}$ describes the effect of the

mutation on the overall stability of the complex [Equation (2)]. Interestingly, if we consider the subset of 216 mutations into alanine that occur at the protein–protein interface, the optimal value of w remains close to 0.7. Accounting for the variation in stability is thus important for the evaluation of the change in binding affinity, even in the case of mutations occurring at the interface.

VALIDATION

SKEMPI data set

We applied our prediction method to a recently published data set, named SKEMPI, in which the experimentally measured effects of mutations on protein–protein binding affinity were compiled (39). This data set is not limited to the results of alanine scanning experiments, and thus includes mutations in any type of amino acid. Out of the 3047 entries in the original data set, we removed 717 multiple mutations, 87 reverse mutations (i.e. from the mutant protein back to the wild type) and 236 redundant entries (i.e. when several experimental values of $\Delta\Delta G_B$ are available for the same mutation in the same protein). In case of redundant entries, the average value of the change in binding free energy was used. The final data set contains 2007 mutations, with $\Delta\Delta G_B$ values ranging from -3.8 to 12.3 kcal/mol. These values are compared with the predictions of BeAtMuSiC on Figure 2.

Strikingly, significant errors in the predictions are observed for a set of 16 mutations of the lysine residue in position 15 of the bovine pancreatic trypsin inhibitor (BPTI), in complex with bovine β-trypsin (BT). Although these mutations severely disrupt the protein–protein interaction, they are predicted to increase or mildly decrease binding affinity (Figure 2). The interaction between BPTI and BT relies mainly on a single residue, Lys 15, which gets inserted into a small cavity on the surface of the protease (Figure 3). The binding affinity is largely determined by the shape complementary of the lysine side-chain and the surface of the cavity, and the highly specific interactions that are created (44). It is not surprising that our coarse-grained approach may sometimes fail to provide an accurate description of the consequences of mutations in such an extreme situation, especially because no modeling and optimization of the conformations of side-chains are performed.

If we discount these 16 mutations, the performances are relatively similar—although somewhat lower—than those observed in the data set of mutations into alanine used to devise our method. The Pearson correlation coefficient between prediction and experiments reaches 0.47 on the full data set and 0.70 on 90% of the data set (Table 1). Note that the 362 mutations into alanine used to identify two parameters of the model are also included in the SKEMPI data set. However, the performances are not affected by the removal of these mutations (correlation coefficient of 0.48 instead of 0.47).

CAPRI experiment

The 26th round of the blind prediction experiment CAPRI (40) was the opportunity to assess the performances of our
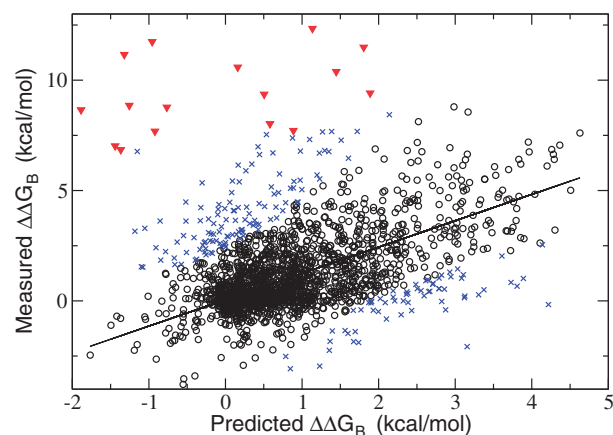
Figure 2. Correlation between predicted and measured changes in binding free energies in the SKEMPI data set. (Black circle) Main data set. (Blue cross) 10% outliers. (Red triangle) Mutations of the lysine at position I15 in the BPTI–BT complex (PDB: 2FTL).
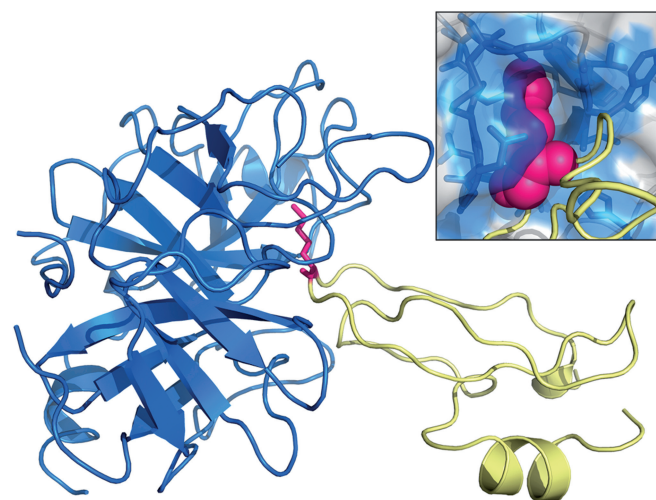
Figure 3. Structure of the complex formed by BPTI (yellow) and bovine BT (blue) (PDB: 2FTL). The lysine residue in position 15 of BPTI is depicted in magenta. The figure was made using PyMOL.

Table 1. BeAtMuSiC performances

Data set	N_{mut}	All mutations		Exclusion of 10% outliers	
		R	σ (kcal/mol)	R	σ (kcal/mol)
Ala scans	362	0.55	1.01	0.76	0.72
SKEMPI	2007	0.40	1.80	0.68	1.19
SKEMPI[a]	1991	0.47	1.59	0.70	1.18

N_{mut} is the number of mutations in the considered data set. R is the Pearson correlation coefficient and σ the root mean square error, between predicted and measured $\Delta\Delta G_B$ values. [a]This data set was obtained after removal of 16 mutations of the Lysine residue at position I15 in the BPTI–BT complex (PDB: 2FTL).

method, in comparison with various approaches developed by other groups. The challenge consisted in the evaluation of the effect of a large number of single-site mutations in two *de novo* designed influenza inhibitors, HB36.4 and HB80.3 (45), on the binding affinity with their target, hemagglutinin. More precisely, predictions were requested for 1007 mutations at 53 positions in the sequence of HB36.4 (target 55), and 855 mutations at 45 positions in the sequence of HB80.3 (target 56).

Out of the 22 participating groups, 18 and 15 submitted predictions for the complete set of mutations in HB36.4 and HB80.3, respectively. The predictions were compared with previously unreleased experimental data concerning the effect of these mutations on the binding properties of the two inhibitors. This data was obtained using deep sequencing of mutant libraries before and after selection for binding (46). Kendall's τ rank correlation coefficient between the predictions of the different groups and the experimental measurements is reported in Figure 4.

The comparison of the results for the two target proteins underlines the difficulty in establishing consistently a detailed ranking of the different approaches. Yet, in both cases, our method stands among the top performers, indicating that the predictive power of BeAtMuSiC is in line with the most efficient state-of-the-art methodologies. The average performance of our method on the two CAPRI targets is somewhat lower than on the other data sets: Kendall's τ reaches 0.36 for the 362 mutations into alanine and 0.29 for the 1991 mutations of the SKEMPI data set. This may be due to the fact that the experimental data used during the CAPRI experiment does not consist in direct measurements of the changes in binding affinity, and that other sources of error may therefore be present.

WEBSERVER

Server input

The main input of the webserver is the structure of the protein–protein complex, in PDB format. The user may either upload his own structure file, or provide the 4-letter PDB code of the structure, which will then be automatically downloaded from the Protein Data Bank (47).

Once the structure has been correctly retrieved, the server will display a summary of the protein chains present in the structure file. The second step consists in the definition of the two partners of the protein–protein interaction. Each chain must be assigned to either the first or second partner, or discarded. Obviously, for the server to provide meaningful predictions, each partner must contain at least one protein chain, and the two partners must be in contact.

The user may then choose to evaluate the effect a few specific mutations, or to perform a systematic scan of all possible mutations in a protein chain or at the protein–protein interface. If several protein chains of identical sequence are present, the mutations will be introduced simultaneously in all of them.
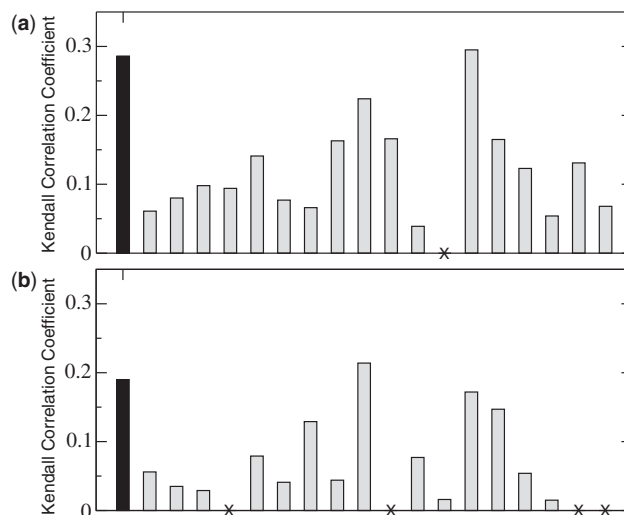
Figure 4. Kendall's τ rank correlation coefficient between predictions and experiments, during the 26th round of the CAPRI experiment (http://www.ebi.ac.uk/msd-srv/capri/round26). The results of our method are depicted in black, and those of other participating methods in gray. Groups that did not submit predictions for the complete set of mutations are not considered here. The symbol 'X' is used when a group submitted a full set of predictions for one target but not for the other. (**a**) Target 55: hemagglutinin-HB36.4. (**b**) Target 56: hemagglutinin-HB80.3. A detailed analysis of the results of this experiment, along with a description of the different prediction methods, will be reported elsewhere (Moretti *et al.*, manuscript submitted).

Server output

The main output of the webserver is the change in binding free energy resulting from each mutation. Users should be aware that when several chains of identical sequence are present, the mutation is introduced simultaneously in each one of those. In such cases, the predictions are not normalized with respect to the number of chains concerned by the mutation, and the reported $\Delta\Delta G_B$ value corresponds thus to the total change in binding free energy between the two selected partners.

The webserver also reports the solvent accessibility of the mutated residue, in the complex and in the individual partners. The solvent accessibility is defined as the ratio of the solvent-accessible surface in the considered structure, as computed by DSSP (48), and in an extended tripeptide Gly-X-Gly (49). BeAtMuSiC identifies a residue as part of the protein–protein interface if its solvent accessibility in the complex is at least 5% lower that in the individual partner. This latter information is provided to the user for convenience, but not used during the computations.

The results may be downloaded as a plain text file, or browsed interactively on the Web site (Figure 5). In particular, if systematic mutations have been performed, the user may choose to display the predictions for mutations at a given position in the sequence, or for mutations with the strongest predicted increase or decrease in binding affinity.
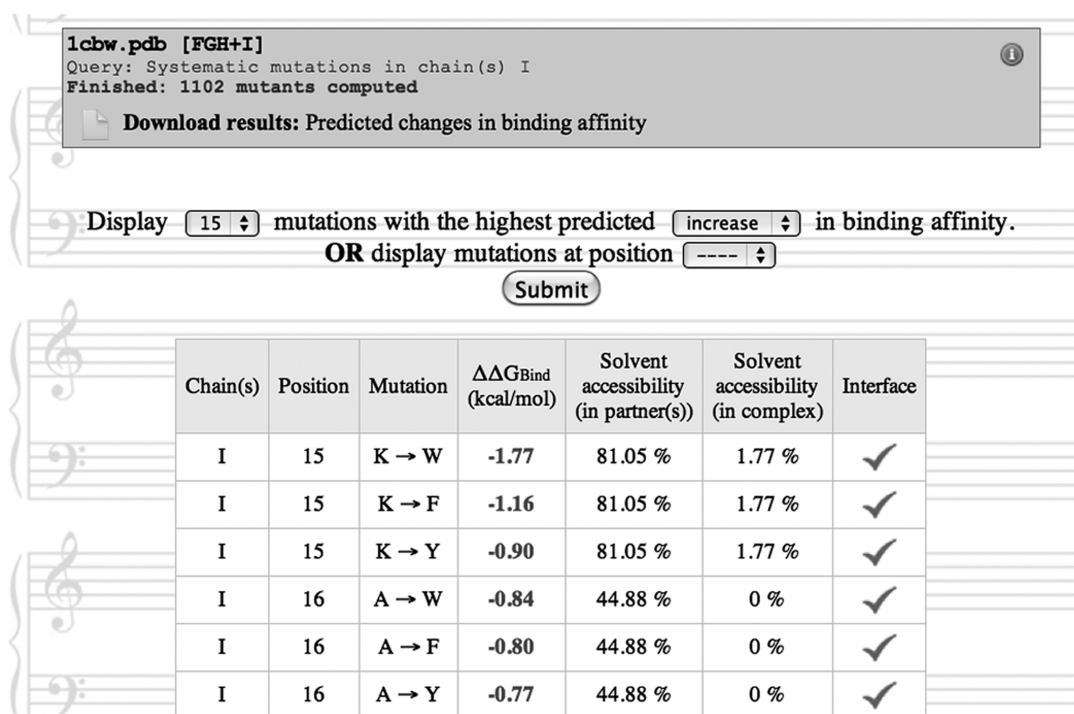
Figure 5. Example output of the BeAtMuSiC server.

DISCUSSION

The prediction of the impact of a mutation on protein–protein binding affinity is more challenging than the prediction of the change in folding free energy. For instance, the method that we previously developed for the latter purpose, on the basis of a model with the same level of coarse-graining, yielded a Pearson correlation coefficient between prediction and experiments of 0.63 on a data set of 2648 mutations, and 0.79 on 90% of the data set (42,50).

In both cases, the accuracy of the predictions ultimately hinges on the quality of the energy function. However, because coarse-grained models do not describe explicitly all of the structural and energetic consequences of a mutation, a number of new obstacles arise when the question of protein–protein binding is considered. Firstly, evaluating mutations occurring outside of the interface region can be challenging, as their impact on binding may be related to an effect on the overall stability of the complex, to an alteration of the dynamical properties and flexibility of one of the interacting partners, or to conformational changes affecting the shape complementarity of the two partners. On the other hand, protein–protein interfaces have been shown to possess distinctive properties from core or surface regions (3–10), and the highly specific nature of many interactions established at those interfaces may be difficult to render accurately without modeling side-chain conformations at atomic resolution. Finally, further investigations would be needed to evaluate whether a single coarse-grained model is sufficient to encompass the large variety of binding modes that characterize protein–protein interactions (individually folded or unfolded partners, transient or permanent interfaces, occurrence of structural rearrangements on binding, etc).

Although these considerations suggest that there is still room for improvement, the results of the 26th round of the CAPRI experiment demonstrated that the predictive power of our method compares well with that of other approaches developed for the same purpose, including predictive models based on a much more detailed structural representation. In addition, the coarse-grained nature of our method provides unique advantages in terms of computational speed, with the possibility to assess rapidly the impact of all possible mutations in a protein chain or at the protein–protein interface. The coarse-graining also ensures that the predictions are robust to imperfections in the structural data. Therefore, a similar level of performance should be expected when using structural models rather than experimentally determined structures (51), provided the relative positioning of the two partners of the interaction is correctly defined.

The BeAtMuSiC server should thus prove useful in a wide range of applications. Typically, protein engineering projects aiming at the design or modification of protein–protein interactions would benefit from the possibility to identify a restricted number of mutations that would constitute ideal candidates for further investigations using more detailed computational approaches and/or experimental tests. Given the importance of binding for the proper biological functioning of many proteins, BeAtMuSiC may also participate to a better understanding of the pathological consequences of some non-synonymous single-nucleotide polymorphisms.

FUNDING

Conflict of interest statement. None declared.

REFERENCES

1. Metz,A., Ciglia,E. and Gohlke,H. (2012) Modulating protein-protein interactions: from structural determinants of binding to druggability prediction to application. *Curr. Pharm. Des.*, **18**, 4630–4647.
2. Karanicolas,J. and Kuhlman,B. (2009) Computational design of affinity and specificity at protein-protein interfaces. *Curr. Opin. Struct. Biol.*, **19**, 458–463.
3. Jones,S. and Thornton,J.M. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
4. Glaser,F., Steinberg,D.M., Vakser,I.A. and Ben-Tal,N. (2001) Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, **43**, 89–102.
5. Chakrabarti,P. and Janin,J. (2002) Dissecting protein-protein recognition sites. *Proteins*, **47**, 334–343.
6. Ofran,Y. and Rost,B. (2003) Analysing six types of protein-protein interfaces. *J. Mol. Biol.*, **325**, 377–387.
7. Reichmann,D., Rahat,O., Cohen,M., Neuvirth,H. and Schreiber,G. (2007) The molecular architecture of protein-protein binding sites. *Curr. Opin. Struct. Biol.*, **17**, 67–76.
8. Bahadur,R.P. and Zacharias,M. (2008) The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cell. Mol. Life Sci.*, **65**, 1059–1072.
9. Gromiha,M.M., Yokota,K. and Fukui,K. (2009) Energy based approach for understanding the recognition mechanism in protein-protein complexes. *Mol. Biosyst.*, **5**, 1779–1786.
10. Berezovsky,I.N. (2011) The diversity of physical forces and mechanisms in intermolecular interactions. *Phys. Biol.*, **8**, 035002.
11. Moreira,I.S., Fernandes,P.A. and Ramos,M.J. (2007) Hot spots–a review of the protein-protein interface determinant amino-acid residues. *Proteins*, **68**, 803–812.
12. Clackson,T. and Wells,J.A. (1995) A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383–386.
13. Morrow,J.K. and Zhang,S. (2012) Computational prediction of protein hot spot residues. *Curr. Pharm. Des.*, **18**, 1255–1265.
14. Wells,J.A. (1991) Systematic mutational analyses of protein-protein interfaces. *Methods Enzymol.*, **202**, 390–411.
15. Thorn,K.S. and Bogan,A.A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284–285.
16. del Sol,A. and O'Meara,P. (2005) Small-world network approach to identify key residues in protein-protein interaction. *Proteins*, **58**, 672–682.
17. Shulman-Peleg,A., Shatsky,M., Nussinov,R. and Wolfson,H.J. (2007) Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol.*, **5**, 43.
18. Ofran,Y. and Rost,B. (2007) Protein-protein interaction hotspots carved into sequences. *PLoS Comput. Biol.*, **3**, e119.
19. Darnell,S.J., LeGault,L. and Mitchell,J.C. (2008) KFC Server: interactive forecasting of protein interaction hot spots. *Nucleic Acids Res.*, **36**, W265–W269.
20. Guney,E., Tuncbag,N., Keskin,O. and Gursoy,A. (2008) HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res.*, **36**, D662–D666.
21. Bromberg,Y. and Rost,B. (2008) Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics*, **24**, i207–i212.
22. Grosdidier,S. and Fernández-Recio,J. (2008) Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics*, **9**, 447.
23. Lise,S., Archambeau,C., Pontil,M. and Jones,D.T. (2009) Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinformatics*, **10**, 365.
24. Cho,K.I., Kim,D. and Lee,D. (2009) A feature-based approach to modeling protein-protein interaction hot spots. *Nucleic Acids Res.*, **37**, 2672–2687.
25. Tuncbag,N., Keskin,O. and Gursoy,A. (2010) HotPoint: hot spot prediction server for protein interfaces. *Nucleic Acids Res.*, **38**, W402–W406.
26. Xia,J.F., Zhao,X.M., Song,J. and Huang,D.S. (2010) APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics*, **11**, 174.
27. Assi,S.A., Tanaka,T., Rabbitts,T.H. and Fernandez-Fuentes,N. (2010) PCRPi: presaging critical residues in protein interfaces, a new computational tool to chart hot spots in protein interfaces. *Nucleic Acids Res.*, **38**, e86.
28. Zhu,X. and Mitchell,J.C. (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins*, **79**, 2671–2683.
29. Wang,L., Liu,Z.P., Zhang,X.S. and Chen,L. (2012) Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Eng. Des. Sel.*, **25**, 119–126.
30. Massova,I. and Kollman,P.A. (1999) Computational alanine scanning to probe protein-protein interactions: a novel approach to evaluate binding free energies. *J. Am. Chem. Soc.*, **121**, 8133–8143.
31. Huo,S., Massova,I. and Kollman,P.A. (2002) Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *J. Comput. Chem.*, **23**, 15–27.
32. Moreira,I.S., Fernandes,P.A. and Ramos,M.J. (2007) Computational alanine scanning mutagenesis–an improved methodological approach. *J. Comput. Chem.*, **28**, 644–654.
33. Bradshaw,R.T., Patel,B.H., Tate,E.W., Leatherbarrow,R.J. and Gould,I.R. (2011) Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction. *Protein Eng. Des. Sel.*, **24**, 197–207.
34. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
35. Kortemme,T. and Baker,D. (2002) A simple physical model for binding energy hot spots in protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
36. Kortemme,T., Kim,D.E. and Baker,D. (2004) Computational alanine scanning of protein-protein interfaces. *Sci. STKE*, **2004**, pl2.
37. Pokala,N. and Handel,T.M. (2005) Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J. Mol. Biol.*, **347**, 203–227.
38. Dehouck,Y., Gilis,D. and Rooman,M. (2012) Design of modified proteins using knowledge-based approaches. *AIP Conf. Proc.*, **1456**, 139–147.
39. Moal,I.H. and Fernández-Recio,J. (2012) SKEMPI: a structural and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–2607.
40. Janin,J., Henrick,K., Eyck,L.T., Sternberg,M.J., Vajda,S., Vakser,I. and Wodak,S.J. (2003) CAPRI: a critical assessment of predicted interactions. *Proteins*, **52**, 2–9.
41. Dehouck,Y., Gilis,D. and Rooman,M. (2006) A new generation of statistical potentials for proteins. *Biophys. J.*, **90**, 4010–4017.
42. Dehouck,Y., Grosfils,A., Folch,B., Gilis,D., Bogaerts,P. and Rooman,M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.
43. Mészáros,B., Simon,I. and Dosztányi,Z. (2011) The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. *Phys. Biol.*, **8**, 035003.
44. Krowarsch,D., Dadlez,M., Buczek,O., Krokoszynska,I., Smalas,A.O. and Otlewski,J. (1999) Interscaffolding additivity: binding of P1 variants of bovine pancreatic trypsin inhibitor to four serine proteases. *J. Mol. Biol.*, **289**, 175–186.

45. Fleishman,S.J., Whitehead,T.A., Ekiert,D.C., Dreyfus,C., Corn,J.E., Strauch,E.M., Wilson,I.A. and Baker,D. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**, 816–821.

46. Whitehead,T.A., Chevalier,A., Song,Y., Dreyfus,C., Fleishman,S.J., De Mattos,C., Myers,C.A., Kamisetty,H., Blair,P., Wilson,I.A. *et al.* (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.*, **30**, 543–548.

47. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

48. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

49. Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985) Hydrophobicity of amino acid residues in globular proteins. *Science*, **229**, 834–838.

50. Dehouck,Y., Kwasigroch,J.M., Gilis,D. and Rooman,M. (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, **12**, 151.

51. Gonnelli,G., Rooman,M. and Dehouck,Y. (2012) Structure-based mutant stability predictions on proteins of unknown structure. *J. Biotechnol.*, **161**, 287–293.